(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification⁷: G01N 31/00, G06F 17/00, 15/00

(21) International Application Number: PCT/US00/20603

(22) International Filing Date: 28 July 2000 (28.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/365,587   30 July 1999 (30.07.1999)   US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US   09/365,587 (CON)
Filed on   30 July 1999 (30.07.1999)

(71) Applicant (for all designated States except US): AGY THERAPEUTICS, INC. [US/US]; Two Corporate Drive, South San Francisco, CA 94080 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): CHIN, Daniel, J. [US/US]; 653 Celestial Lane, Foster City, CA 94404 (US). HENDRIX, Donna [US/US]; 1318 Milvia Street, Berkeley, CA 94709 (US). ZHAO, Oliver [CA/US]; 25 Poncetta #220, Daly City, CA 94015 (US).

(74) Agents: APPLE, Randolph, T. et al.; Townsend and Townsend and Crew LLP, 8th Floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR,

[Continued on next page]

(54) Title: TECHNIQUES FOR FACILITATING IDENTIFICATION OF CANDIDATE GENES
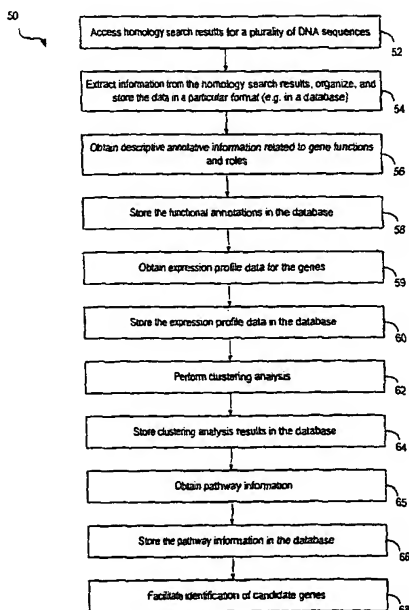
(57) Abstract: Techniques for facilitating the identification of candidate genes from a plurality of DNA sequences. According to an embodiment of the present invention, techniques are provided for extracting and integrating information from various information sources and results of various analyses, and storing the integrated information in a form which is conducive to identification of candidate genes. The stored information may include results of a homology search for the plurality of DNA sequences, annotative information for the plurality of DNA sequences indicating the biochemical functions and physiological roles of the plurality of DNA sequences, gene expression profile data for the plurality of DNA sequences describing behavioral patterns of the plurality of DNA sequences, results from clustering the plurality of DNA sequences based on time course data as described by the gene expression profile data, and other information.

WO 01/13105 A1

LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*
— *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

TECHNIQUES FOR FACILITATING IDENTIFICATION OF CANDIDATE GENES

COPYRIGHT NOTICE

5   A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the xerographic reproduction by anyone of the patent document or the patent disclosure in exactly the form it appears in the U.S. Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

10

CROSS-REFERENCES TO RELATED APPLICATIONS

          This application also claims priority from and is a continuation-in-part application of non-provisional U.S. Patent Application No. 09/365,587, entitled "SYSTEM AND METHOD FOR IDENTIFYING CRITICAL REGULATED GENES"

15   filed July 30, 1999, the entire contents of which are herein incorporated by reference in their entirety for all purposes.

BACKGROUND OF THE INVENTION

          The present invention relates generally to the field of bioinformatics, and

20   more particularly to techniques for facilitating the identification of candidate genes.

          With recent advances in the identification of expressed sequence tags (ESTs) and the sequencing of the human genome, a number of researchers are now directing their efforts towards analyzing the data from the genome maps and sequences. A significant portion of this research is being directed towards identifying genes which

25   might trigger, prevent, ameliorate, or somehow affect a variety of diseases or physiological states. Such genes are commonly referred to as "candidate" genes.

          The identification of candidate genes is critical to entities such as drug companies who may use the information related to the candidate genes to identify better drug targets in the drug development process. The early identification of candidate genes

30   could reduce the number of potential therapeutics moving through a company's clinical testing pipeline, significantly reducing overall costs and reducing the time taken by the company to market the drugs.

However, conventional techniques do not facilitate easy identification of candidate genes. This is due to the enormous amount of information being generated by the researchers, and the lack of adequate tools to organize the information in a manner which facilitates analysis of the information. For example, techniques such as parallel

5   expression and analysis using cDNA arrays, as described in U.S. Patent No. 5,807,522, and synthetic DNA array technology, as described in U.S. Patent No. 5,593,839 and U.S. Patent No. 5,571,639, have been developed to study large scale gene expression profiles (e.g. time-courses of a disease process or comparisons between an altered physiologic or metabolic state with an untreated biological sample). Databases and algorithms have also

10 been developed to analyze the results of the above-mentioned array technologies. Public databases of metabolic, genetic and physiological pathways of yeast (e.g., Munich Information Center for Protein Sequences (MIPS)) and some mammalian genes (e.g., Kyoto Encyclopedia of Genes and Genomes (KEGG)) have been developed largely from the published literature of many traditional low-throughput experimental studies.

15 However, the information provided by the various sources of information identified above and other sources has not been integrated in a coherent manner conducive to identification of candidate genes.

Based on the foregoing, there is a need for techniques which can facilitate the identification of candidate genes. It is desirable that these techniques be able to

20 correlate various types of information and store it in a format which can be easily accessed or queried by researchers interested in identifying candidate genes.

SUMMARY OF THE INVENTION

The present invention discusses techniques for facilitating identification of

25 candidate genes from a plurality of DNA sequences. According to an aspect of the present invention, techniques are provided for extracting and integrating information from various information sources and results of various analyses, and storing the integrated information in a form which facilitates identification of candidate genes.

According to an embodiment, the present invention accesses results of a

30 homology search for the plurality of DNA sequences, annotative information for the plurality of DNA sequences indicating the biochemical functions and physiological roles of the plurality of DNA sequences, gene expression profile data for the plurality of DNA sequences describing behavioral patterns of the plurality of DNA sequences, results from clustering the plurality of DNA sequences based on the behavioral patterns of the

plurality of DNA sequences as described by the gene expression profile data, and other information. The information accessed by the present invention is stored in a format, e.g. a database, which facilitates identification of candidate genes.

According to another embodiment, the present invention receives queries

5   identifying criteria for the candidate genes. In response to the queries, the present invention searching the database storing information for the plurality of DNA sequences to identify a set of DNA sequences which satisfy the query criteria. The set of DNA sequences are then output as a result of the query.

According to yet another embodiment of the present invention, a user may

10   configuring a query identifying criteria for the candidate genes and communicate the query to a server storing information related to a plurality of DNA sequences. According to this embodiment, the information related to the plurality of DNA sequences may comprise results of a homology search for the plurality of DNA sequences, annotative information for the plurality of DNA sequences describing the biochemical functions and

15   physiological roles of the plurality of DNA sequences, gene expression profile data for the plurality of DNA sequences describing behavioral patterns of the plurality of DNA sequences, results from clustering the plurality of DNA sequences based on the behavioral patterns of the plurality of DNA sequences as described by the gene expression profile data, and other information. In response to the query, the user receives

20   a first set of DNA sequences which satisfy the criteria for the candidate genes identified in the query.

The invention will be better understood by reference to the following detailed description and the accompanying figures.

25   BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a simplified block diagram of a distributed computer network incorporating an embodiment of the present invention;

Fig. 2 is a simplified block diagram of a computer system according to an embodiment of the present invention;

30   Fig. 3 is a simplified flowchart showing processing performed by an embodiment of the present invention to facilitate identification of candidate genes from a plurality of input DNA sequences;

Fig. 4 depicts a process of performing homology analysis for a plurality of sequences according to an embodiment of the present invention;

3

Fig. 5 depicts a database schema showing information extracted from homology search results and stored for the query cDNA sequences according to an embodiment of the present invention;

Fig. 6 is a simplified flowchart showing processing performed by an
5   embodiment of the present invention for obtaining descriptive annotative information for the genes;

Fig. 7 depicts a database schema showing the functional annotative information stored for the genes according to an embodiment of the present invention;

Fig. 8 depicts a database schema showing the gene expression profile data
10 stored for the genes according to an embodiment of the present invention; and

Fig. 9 is an exemplary look-up table for general rankings of biomedical journals.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

15   The present invention discusses techniques for facilitating identification of candidate genes from a plurality of DNA sequences. According to an aspect of the present invention, techniques are provided for extracting and integrating information from various information sources and results of various analyses, and storing the integrated information in a form which facilitates identification of candidate genes.

20   As part of the analysis, an embodiment of the present invention analyzes and extracts information from homology searches performed on a plurality of DNA sequences. According to another aspect, the present invention extracts descriptive annotative information from various information stores about cDNA clones, which have been isolated on the basis of differential expression from various disease models or
25 altered physiological states. According to another embodiment, the present invention extracts information about causally ordered (i.e. as defined by autoregression-based causality analysis) behavioral patterns of differentially expressed cDNAs from gene expression profile data. According to another embodiment, the present invention correlates the descriptive annotative information about cDNA clones with numerical
30 experimental data on the behavior of the cDNAs extracted from, for example, the gene expression profiling data. According to another embodiment, the present invention integrates the information to provide a model to facilitate experimental testing of the candidate genes. The information extracted/obtained by the present invention is stored in

4

a database. According to an embodiment of the present invention, users may query the information stored in the database to identify candidate genes.

Fig. 1 is a simplified block diagram of a distributed computer network 10 incorporating an embodiment of the present invention. Computer network 10 includes a number of client systems 16-1, 16-2, and 16-3, and a server system 14 coupled to a communication network 12 via a plurality of communication links 18. Communication network 12 provides a mechanism for allowing the various components of distributed network 10 to communicate and exchange information with each other. Communication network 12 may itself be comprised of many interconnected computer systems and communication links. Communication links 18 may be hardwire links, optical links, satellite or other wireless communications links, wave propagation links, or any other mechanisms for communication of information. While in one embodiment, communication network 12 is the Internet, in other embodiments, communication network 12 may be any suitable computer network. Distributed computer network 10 depicted in Fig. 1 is merely illustrative of an embodiment incorporating the present invention and does not limit the scope of the invention as recited in the claims. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. For example, more than one server system 14 may be coupled to communication network 12.

Client systems 16 typically request information from a server computer system which provides the information. For this reason, servers typically have more computing and storage capacity than client systems. However, a particular computer system may act as both as a client or a server depending on whether the computer system is requesting or providing information. Additionally, although the invention has been described using a client-server environment, it should be apparent that the invention may also be embodied in a stand-alone computer system.

According to the teachings of the present invention, server system 14 is responsible for obtaining and storing information for a plurality of DNA sequences in order to facilitate identification of candidate genes from the DNA sequences. Server system 14 may store the information in one or more databases accessible to server 14. These databases may be locally coupled to server 14 or may be distributed across distributed computer network 10 and accessed by server 14 via communication network 12.

5

Software modules executing on server system 14 are responsible for obtaining information from a plurality of information sources, and integrating and storing the information in a manner which facilitates identification of candidate genes. The information sources may include databases accessible to server system 14, results from

5   various analyses, published sources of information such as magazine articles, etc., and other like information sources. Server 14 also provides services allowing users to select, access, retrieve, or query information stored by the server.

Server 14 is responsible for receiving information requests from client systems 16, performing processing required to satisfy the requests, and for forwarding the

10   results corresponding to the requests back to the requesting client system. The processing required to satisfy the request may be performed by server 14 or may alternatively be delegated to other servers connected to communication network 12.

According to the teachings of the present invention, client systems 16 enable users to access and query information stored by server system 14. In a specific

15   embodiment, a "web browser" application executing on a client system enables users to select, access, retrieve, or query information stored by server system 14. Examples of web browsers include the Internet Explorer browser program provided by Microsoft Corporation, and the Netscape Navigator browser provided by Netscape Corporation, and others.

20   Fig. 2 is a simplified block diagram of computer system 20 according to an embodiment of the present invention. Computer system 20 typically includes at least one processor 24 which communicates with a number of peripheral devices via bus subsystem 22. These peripheral devices typically include a storage subsystem 32, comprising a memory subsystem 34 and a file storage subsystem 40, user interface input devices 30,

25   user interface output devices 28, and a network interface subsystem 26. The input and output devices allow user interaction with computer system 20. It should be apparent that the user may be a human user, a device, another computer, and the like. Network interface subsystem 26 provides an interface to outside networks, including an interface to communication network 12, and is coupled via communication network 12 to

30   corresponding interface devices in other computer systems.

User interface input devices 30 may include a keyboard, pointing devices such as a mouse, trackball, touchpad, or graphics tablet, a scanner, a touchscreen incorporated into the display, audio input devices such as voice recognition systems, microphones, and other types of input devices. In general, use of the term "input device"

6

is intended to include all possible types of devices and ways to input information into computer system 20 or onto computer network 12.

User interface output devices 28 may include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices. The display
5   subsystem may be a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), or a projection device. The display subsystem may also provide non-visual display such as via audio output devices. In general, use of the term "output device" is intended to include all possible types of devices and ways to output information from computer system 20 to a human or to another machine or computer
10   system.

Storage subsystem 32 stores the basic programming and data constructs that provide the functionality of the various systems embodying the present invention. For example, the various modules implementing the functionality of the present invention may be stored in storage subsystem 32. These software modules are generally executed
15   by processor 24. In a distributed environment, the software modules may be stored on a plurality of computer systems and executed by processors of the plurality of computer systems. Storage subsystem 32 also provides a repository for storing the various databases storing information according to the present invention. Storage subsystem 32 typically comprises memory subsystem 34 and file storage subsystem 40.

20   Memory subsystem 34 typically includes a number of memories including a main random access memory (RAM) 38 for storage of instructions and data during program execution and a read only memory (ROM) 36 in which fixed instructions are stored. File storage subsystem 40 provides persistent (non-volatile) storage for program and data files, and may include a hard disk drive, a floppy disk drive along with
25   associated removable media, a Compact Digital Read Only Memory (CD-ROM) drive, an optical drive, removable media cartridges, and other like storage media. One or more of the drives may be located at remote locations on other connected computers at another site on communication network 12. Information stored according to the teachings of the present invention may also be stored by file storage subsystem 40.

30   Bus subsystem 22 provides a mechanism for letting the various components and subsystems of computer system 20 communicate with each other as intended. The various subsystems and components of computer system 20 need not be at the same physical location but may be distributed at various locations within distributed

7

network 10. Although bus subsystem 22 is shown schematically as a single bus, alternate
embodiments of the bus subsystem may utilize multiple busses.

Computer system 20 itself can be of varying types including a personal
computer, a portable computer, a workstation, a computer terminal, a network computer,
5   a television, a mainframe, or any other data processing system. Due to the ever-changing
nature of computers and networks, the description of computer system 20 depicted in Fig.
2 is intended only as a specific example for purposes of illustrating the preferred
embodiment of the present invention. Many other configurations of a computer system
are possible having more or less components than the computer system depicted in Fig. 2.
10 Client computer systems 16 and server computer systems 14 generally have the same
configuration as shown in Fig. 2, with the server systems generally having more storage
capacity and computing power than the client systems.

Fig. 3 depicts a simplified flowchart 50 showing processing performed by
an embodiment of the present invention to facilitate identification of candidate genes
15 from a plurality of input DNA sequences. As shown in Fig. 3, processing is initiated
when the server system 14 accesses results of a homology search from the plurality of
input DNA sequences (step 52).

The DNA sequences which are input as queries to the homology search are
generally complementary DNA (cDNA) sequences which have been synthesized using
20 isolated messenger RNA (mRNA) sequences, which are the transcription products of
expressed genes. The cDNA sequences are used as input sequences to the homology
search analysis since cDNAs represent expressed genomic regions and are thus believed
to identify parts of the genome with the most biological and medical significance.

As part of the homology search, DNA and protein sequence databases are
25 searched to find sequences which are related to the input or query DNA sequences. For
example, given a set of differentially expressed query cDNA sequences corresponding to
the mRNA of their cognate genes, a homology search identifies known, similar and
unknown genes. A homology search is generally performed by using computer-
implemented search algorithms to compare the query cDNA sequences with sequence
30 information stored in a plurality of databases accessible via a communication network, for
example, the Internet. Examples of such algorithms include the Basic Local Alignment
Search Tool (BLAST) algorithm, the PSI-blast algorithm, the Smith-Waterman algorithm,
the Hidden Markov Model (HMM) algorithm, and other like algorithms. For example, a
"blastn" program utilizing the BLAST algorithm may be used to search the Genbank

8

database for homologs of the query cDNA sequences. According to an embodiment of the homology search, the query cDNA sequences may be grouped as "known," "unknown," or "similar" sequences. "Known" cDNA sequences include sequences with substantial sequence identity to existing sequence entries in a sequence database, such as

5   the GenBank database. "Unknown" cDNA sequences include sequences similar to existing sequence entries in a sequence database but lacking functional annotation, or those sequences with no matching sequences in existing sequence databases. "Similar" cDNA sequences include sequences for which no matches are found in the sequence database, but which exhibit similarity, as defined below, to existing entries in sequence

10   databases.

Two or more sequences may exhibit "substantial sequence identity" if the sequences have at least 70%, preferably 80%, most preferably 90%, 95%, 98% or 99% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using a particular sequence comparison algorithm or by

15   using visual inspection.

Several different sequence comparison techniques may be used. According to a first technique, two sequences (amino acid or nucleotide) can be compared over their full-length (e.g. the length of the shorter of the two, if they are of substantially different lengths) or over sub-sequences of at least 200, about 200, about 500 or about

20   1000 contiguous nucleotides or at least about 40, about 50, or about 100 contiguous amino acid residues. According to an embodiment of the present invention, a query cDNA sequence may qualified as a "known" gene if the query DNA sequence meets the following stringent criteria: (1) a sequence length greater than 200 nucleotides with greater than or equal to 80% identity over 70% of the query sequence length with an E-

25   value (a probability value of a match occurring if the sequence were randomized) of less than 1e-50; and (2) for the predicted amino acid homology, greater than or equal to 80% identity for a segment length greater than 50 amino acids and an E-value of less than 1e-20. Sequences that meet either, but not both, the DNA or protein sequence criteria may be grouped as "similar" genes after examination of the respective DNA or protein

30   alignments.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input to a computer, subsequence coordinates

9

are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

5   As stated above, a plurality of homology search algorithms may be used to determine optimal alignment of sequences. These include the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482 (1981), the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol., 48:443 (1970), the similarity method of Pearson & Lipman, Proc. Natl. Acad. Sci. USA 85:2444 (1988), the PSI-Blast

10   homology algorithm of Altschul et al., Nucleic Acids Res. 25:3389-402 (1997), the computerized implementations of algorithms GAP, BESTFIT, FASTA, and TFASTA included in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), by Hidden Markov Models (HMM, Durbin, Eddy, Krogh & Mitchison, Cambridge University Press, 1998), or EMotif/EMatrix to identify sequence

15   motifs (Nevill-Manning, Wu, & Brutlag, Proc Natl. Acad. Sci U S A. 1998 May 26;95(11):5865-71), or by visual inspection (see generally Ausubel et al., supra). Each of the above identified algorithms and the references are herein incorporated by reference in its entirety for all purposes. These algorithms are well known to one of ordinary skill in the art of molecular biology and bioinformatics. When using any of the aforementioned

20   algorithms, the default parameters for "Window", gap penalty, etc., are used. Practitioners of the art molecular biology with average skill will recognize these parameters as: (a) the "window" is typically a 9, 10 or 11 nucleotide word length of sequence over which the homology is determined; and (b) gap penalty is a scoring value to prevent large gaps from occurring in reported alignments.

25   The BLAST algorithm is well suited for determining percent sequence identity and sequence similarity. The BLAST algorithm is described in Altschul et al., J Mol. 215:403-410, (1990), the entire contents of which are herein incorporated by reference for all purposes. Several software programs incorporating the BLAST algorithm are publicly available through the National Center for Biotechnology

30   Information (NCBI) (http://www.ncbi.nlm.nih.gov/). These programs include the blastp, blastn, blastx, tblastn, tblastx, and PSI-blast software programs. Due to codon wobble or species differences, more informative homologies can be found by comparing the

10

predicted protein sequence of a DNA query sequence to a protein sequence database. For this task, the Smith-Waterman or PSI-BLAST algorithms may be used. Similarly, for weak homologs, functional domains of proteins may be discerned by Smith-Waterman, HMM or Emotif algorithms. Software for performing HMM and Smith-Waterman

5 analysis can be obtained from a variety of public sources (e.g. http://hmmer.wustl.edu/; http://www.stanford.edu/~sntaylor/bioc218/final.htm#Appendix) and/or from vendors that sell accelerated computer hardware to rapidly process large batches of sequences (e.g. Paracel, Pasadena, CA or Time-Logic, Reno, NV). Software for EMotif/Ematrix can be obtained from sources such as the Brutlag Bioinformatics Group, Stanford

10 University, Stanford, CA.

The BLAST heuristic search algorithm is optimized for speed and searches sequence databases accessible to server 14 for optimal local alignments to the input query DNA sequences. Databases which may be searched using the BLAST programs include the SWISS-PROT protein sequence database, GenBank database, the Genome Sequence

15 database (GSDB), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence database, the DNA Database of Japan (DDBJ), and other like databases.

The BLAST algorithm identifies high scoring sequence pairs (HSPs) by identifying short words of length "W" in the query cDNA sequence, which either match or satisfy some positive-value threshold score "T" when aligned with a word of the same

20 length in a database sequence. "T" is referred to as the neighborhood word score threshold (Altschul et al, supra). An "X" parameter is a positive integer representing the maximum permissible decay of the cumulative segment score during word hit extension. These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each

25 sequence for as far as the cumulative alignment score can be increased. Extension of the word hits in each direction are halted when the cumulative alignment score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments, or when the end of either sequence is reached. The BLAST algorithm parameters "W", "T", and "X" determine the sensitivity and speed of the alignment. Accordingly, the

30 stringency of a BLAST search can be adjusted by appropriately setting the search parameters. However, if the search parameters are too loose, an excessive amount of biologically questionable "hits" may be returned. The BLAST program uses as defaults a

11

wordlength (W) of 11, the BLOSUM62 scoring matrix (see Henikoff & Henikoff, Proc. Natl. Acad. Sci. USA 89:10915 (1989)) alignments (B) of 50, expectation (E) of 10, M=5, N-4, and a comparison of both strands. Typically, the default parameters can yield from zero to scores of likely homologs for the input query DNA sequences.

5   In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g. Karlin & Altschul, Proc. Natl. Acad. Sci. USA 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N) or E-value as an expected value), which provides an indication of the probability by which a

10   match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.01, more preferably less than about 0.001, and most preferably less than about 0.0001.

15   A further indication that two nucleic acid sequences or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions. These

20   polypeptide sequence comparisons are enabled by the Smith-Waterman, HMM and EMotif algorithms.

           As is well known to one of ordinary skill in the art, results from a homology search or analysis includes: a plurality of cDNA query sequences; a list of homologous (target) sequences; an E-Value that describes the probability that the original

25   (query) sequence match with the target sequence could occur randomly; the annotation of the target sequence, if provided; an alignment of the query sequence to each target sequence; the percent identity of the query sequence to the target sequence; the hit length, or length of the sequence over which the percent identity is determined.

           The complete homology analysis of a plurality of sequences according to

30   an embodiment of the present invention is composed of a process described in Fig. 4. The output(s) from the process shown in Fig. 4 may be used as the input to step 52 in Fig. 3. The rationale for this sequential strategy of homology analysis is to automate the method of sequence classification. According to the embodiment shown in Fig. 4, input

12

sequences 80 are subjected to BLAST analysis 82 against an internal database of cDNA sequences 84. Near identical homologs (E-value < 1e-80) are sieved and recorded as being strong homologs of previously classified entries 86 of the internal database. Those sequences failing this test, are subjected to blastn analysis 88 against the GenBank

5   nucleotide (NT) and patent databases 90. Those sequences showing strong similarity (E-value < 1e-50 with sequence length > 200 nucleotides, 80% identity over 70% of the query sequence length) are classified as "known" genes 92. Those sequences failing this test are subjected to Smith-Waterman analysis 94 against the protein databases of Swiss-Prot and the translated patent database 96. Those sequences with E-values < 1e-20 with

10   80% identity over a segment length > 50 amino acids are classified as "known" genes 98 while sequences with an E-value > 1e-20 are subjected in parallel to (a) HMM 102 and EMotif 100 analysis against the Swiss-Prot and GenBank non-redundant (NR) protein databases 104 and (b) BLASTN analysis 106 against the GenBank EST and genomic databases 108. Those sequences with an E-value <1e-9 after HMM or EMotif are scored

15   as "Similar" genes 110 while sequences with an E-value < 1e-60 after the final BLASTN analysis 106 are classified as "unknown" 112. Any sequences failing this last test, are classified as "Novel" 114.

The present invention extracts relevant information from the homology analysis output as described above for each input DNA sequence, organizes the

20   information, and stores it in a format which facilitates further processing and analysis of the information (step 54). According to an embodiment of the present invention, the information extracted from the BLAST, Smith-Waterman and HMM search output is stored in a database. The information extracted and stored by the present invention during step 54 is shown by the database schema depicted in Fig. 5. Figs. 7 and 8 depict

25   other database structures for storing information according to an embodiment of the present invention.

Fig. 5 shows information (database table "HomologyResults" 120) which is extracted from the homology search results, and stored for each query cDNA sequence according to an embodiment of the present invention. It is important to note that multiple

30   (typically 10) homologs for each query sequence are stored in this database table in order to facilitate extraction of the most descriptive and accurate annotation for the query sequence. It should also be evident that various other formats, in addition to tables and databases, may also be used to store the information. The following scenario is common: the top 1, 2, 3, 4 or 5 blastn homologs of a query have E-values within a 10-fold range

13

and are < 1e-50 yet lack informative annotative information (e.g. such homologs are expressed sequence tags or genomic DNA). However, the second, third, fourth, fifth, sixth or seventh homolog's E-values might have the following attributes: the E-value is less than 1e-50 and is within 10 or 100 fold of the top hit but the weaker homolog's

5   annotation might provide more informative description of the query sequence's role or function; e.g. the weaker homolog might be an enzyme, receptor or structural protein. Identification of these more accurate descriptions are facilitated by a combination of keyword tables and information extraction methods described herein. In these circumstances, those of normal skill in the art of bioinformatics will recognize that the

10   weaker hit provides the most useful annotation provided the E-value meets the above criteria.

     For each homolog, the present invention stores, in database tables "DNAsequence" 130 and "HomologyResults" 120, the name of the sequence (attribute "seqFile" 130-a and 120-a), the sequence ("Sequence" 130-b), the quality scores or Phred

15   values (Ewing, Hiller, Wendl & Green, Genome Research, 8:175-185, 1998), ("QualityScores" 130-c), the accession number of any homolog, i.e. the GenBank identifier number ("GID"120-e), the best GID derived from BLAST analysis ("BestBlastnGID" 130-f), the best GID derived from BLAST against the patent DNA database analysis ("BestPatent-GID" 130-g), the best GID derived from Smith-Waterman

20   analysis derived from the Swiss-Prot database ("BestSW-GID" 130-h), the best GID derived from Smith-Waterman analysis of the patent (database "BestPatent-SW-GID" 130-i), the best GID derived from the best human homolog in BLAST analysis ("BestHumanBlastn-GID" 130-j), and the best GID derived from the best human homolog derived from Smith-Waterman analysis ("BestHuman-SW-GID" 130-k). For

25   any homolog, the algorithm (e.g. BLAST or HMM) used for the homology search is recorded ("Algorithm" 120-b), the frame of the predicted protein for protein comparisons ("Frame" 120-c), the database searched ("Database" 120-d), the GenBank annotation for any homolog ("HitDescription" 120-f), the species of the annotation ("Species" 120-g), the E-value ("E-value" 120-h), the length of the alignment region ("AlignLength" 120-i),

30   the percent identity of the aligned sequences ("PercentIdentity" 120-j), the length of the query in the alignment ("QueryLength" 120-k), the length of the target in the alignment ("TargetLength" 120-l), a number representing the fraction of the total query length represented in the hit region ("ALength/QLength" 120-m), the start position of the query sequence in the alignment ("QueryStart" 120-n), the position of the end of the query

14

("QueryEnd" 120-o), the start position of the target sequence ("TargetStart" 120-p), the end position of the target sequence ("TargetEnd" 120-q), the query sequence in the alignment ("QSequence" 120-r), the consensus of the alignment ("Consensus" Fig. 120-s), and the target sequence in the alignment ("TSequence" 120-t).

5   Referring back to Fig. 3, server 14 then obtains (step 56) descriptive annotative information on the biochemical function(s) and the physiological role(s) for the known genes from the plurality of cDNA sequences and stores the information in the database (step 58). Fig. 6 depicts a simplified flowchart 140 showing processing performed by an embodiment of the present invention for obtaining descriptive annotative

10   information for the known genes. As shown in Fig. 6, several different techniques may be used by the present invention to obtain the functional information. According to a first technique, the present invention accesses information sources containing functional information related to the known genes (step 142). The information sources may include articles, published material, and other like material accessible to server 14. According to

15   a specific embodiment, the present invention may use the accession numbers or the GenBank identifiers (GIDs) associated with the DNA sequences and their homologs to find the published material. Text processing tools may then be used by the present invention to automatically extract functional information from the information sources accessed in step 142 (step 146). The extracted information may then be summarized (step

20   148) and stored in the database (step 150).

According to another technique, the present invention may obtain the functional information from databases storing functional information and which are accessible to server 14 (step 144). Examples of such databases include databases provided by Proteome of Boston, Massachusetts, DoubleTwist of Oakland, California, the

25   Genbank database of deposited DNA and protein sequence data (http://www.ncbi.nlm.nih.gov:80/entrez/), the SWISS-PROT protein database (http://www.expasy.ch/sprot/), the PubMed or Medline (NCBI) (http://www.ncbi.nlm.nih.gov) databases of abstracts derived from thousands of peer-reviewed biomedical journals, and other like databases. The Proteome databases are

30   concise descriptions of known genes, their protein products and their functions and roles and known interactors as described in the current literature. The information extracted from the published material and genomic databases may then be summarized (step 148) and stored in the database (step 150).

15

The GenBank record of a cDNA or gene sequence commonly contains references to peer-reviewed publication information, stored in the Medline database about the gene. The Medline database can be accessed via the Internet via the PubMed interface (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi). Alternatively, the

5   GenBank record contains informative keywords related to the gene which may be used to perform broad topic searches on the Medline database. For example, protein products of genes participate in many processes essential to metabolism, development and reproduction. In some cases, a protein encoded by a gene may have more than one function and/or more than one role. For example, the yeast inositol 1-4-5 triphosphate

10   kinase enzyme adds a phosphate moiety to phosphoinositol– an important component involved in signaling. However, this protein also can act as a regulatory scaffolding protein for transcription factors in the nucleus (Audrey R. et al. Science 287:2026-2029, 2000). Thus, this single protein can function as both an enzyme and a structural protein. Similarly, this gene product has two roles: it can participate in signaling processes and

15   mRNA transcription. These instances are also examples of general pathways but further annotative information from the published literature could refine these topics to even more specific pathways. For example, the enzymatic activity might be most important for a growth hormone pathway and the structural role might be more important to a specific subset of transcription factors engaged in controlling cell division. In this invention,

20   these relational links between genes and cellular or organismal processes constitute a web of interacting pathways that are extracted accurately and comprehensively.

The biological demands for information extraction from published material, such as abstracts, etc., in a comprehensive and consistent manner is unique to the world of manual biological annotation. Traditionally, extraction of information was

25   done manually with varying degrees of consistency and accuracy. With recent advances in information extraction technologies, various software programs have been developed to automate information extraction and to summarize the extracted information. Examples of such programs include programs provided by Inxight Corp. of Santa Clara, California. Another example of a software package for information or knowledge

30   extraction is the Crystal-Badger-Marmot suite from the Center for Intelligent Information Retrieval, Univ. of Massachusetts, Amherst, MA. Such software programs have been applied to extract information from abstracts of published papers as well as from full-text

16

papers. According to an embodiment of the present invention, these techniques are applied to generate tables of genes, tables of pathways composed of genes, and tables of relationships between and amongst genes and pathways. As described below, the validation of a relationship between or amongst genes is evaluated in a quantitative

5   fashion.

According to an embodiment of the present invention, information extraction programs, such as those discussed above and others, may be used to extract (step 146 in Fig. 6) descriptive annotation information from information accessible to server 14 and to summarize (step 148 in Fig. 6) the information. According to an aspect

10   of the present invention, the annotative information is stored in a database.

According to the present invention, information is extracted and stored for both the majority views and potentially multiple minority views. This is due to dramatic shifts in the understanding of biological systems over time. These shifts are also referred to as "paradigm shifts" (Kuhn, T., The Structure of Scientific Revolutions, Univ. Chicago

15   Press 1962). According to these paradigm shifts, a minority view becomes accepted as being the correct interpretation after critical new data is acquired. The change in accepted "truth" of a paradigm can be dramatic or subtle in various domains of knowledge, and in the realm of biology both extremes can occur– hence the need for comprehensive collections of entity-relationships amongst genes, functions, roles and pathways. The

20   need for storing both the majority and minority views becomes important when one realizes that the laws of biology are not yet deterministically known. This is substantially different from prior art bioinformatics techniques which only stored information related to the majority view (e.g. T. Rindflesch, L. Tanabe, J. Weinstein & L. Hunter PSB 2000:517-528).

25   For example, for a given biological topic, perhaps 51, 75, or 90 out of 100 published abstracts may describe a phenomenon as being caused by the interactions of genes A and B whereas a smaller subset of abstracts, perhaps 10, 25 or 49 may describe a more complex interaction between genes A and C prior to gene B. The former A-B model would be considered the consensus, "majority view" model (a "truth") and the

30   latter A-C-B model would be considered a "minority view" and likely regarded as being "false." According to traditional bioinformatics techniques, only information related to

17

strict "truths" was maintained and information related to the minority view(s) was discarded to reduce the amount of data being stored.

According to an embodiment of the present invention, minority views (e.g. unusual or unexpected relationships between genes or metabolic pathways) are also
5   stored in the database but assigned a lower reference score (see "RefScore " attribute 200-k in table "Reference" 200 in Fig. 7, "FunctionScores" attribute 170-g in table "Function" 170, "RoleScores" attribute 180-1 in table "Role" 180, and attributes 220-a through 220-f of table "RefScore" 220.) associated with the descriptive annotation of the known genes from the plurality of cDNA. The reference score (or their summary scores,
10   "FunctionScores" 170-g and "RoleScores" 180-1) quantizes the "acceptance/majority opinion" for an alleged role or function of a gene. Of particular importance to "minority" views is the extraction and recording of special circumstances or boundary conditions under which the phenomena or relationship amongst genes might exist. For example, information related to minority views (e.g. unusual or unexpected relationships between
15   genes or metabolic pathways) is stored in the database but assigned a lower reference score than information associated with the majority view. The metric for evaluating a specific published reference article also assigns a score derived from the Citation Index database (Institute for Science Information, Philadelphia) which quantitatively ranks the impact of a given paper by the number of times that paper is subsequently referenced.
20   For the most significant papers, a published article can be referenced thousands of times. The Citation Index also ranks journals with high impact but only from the same criteria of frequently-cited papers from the journals regardless of whether the published paper is ultimately revised or shown to be inaccurate or limited to a set of conditions. Hence, one embodiment of this invention provides a mechanism to take into account the quality of
25   the information source. This is both general and a specific measure. In general, articles in journals respected by a consensus of biomedical and genomics practitioners are believed to be reliable. For example, a publication in journals with a recognized, rigorous peer-review process (e.g. Science, Nature, the Journal of Biological Chemistry, or the Journal of Clinical Investigations) would receive 100 points or > 90 points whereas
30   publication in "lesser" journals (e.g. Journal of Antisense Research or Experimental Cell Research) would only receive 10 or 40 points.

18

Fig. 9 is an exemplary look-up table for general rankings of such biomedical journals. However, scores from Fig. 9 may be adjusted because the information source's peer-review process can be dependent upon the reviewers for a given domain or the degree of democratic consensus of a journal's editorial board. A

5   domain specific weighting factor is derived for the major journals and can be applied systematically while in other cases, a human annotator must make the judgment. The adjustment can range between 10 and 50% of the original score and an article in a "lower-quality" journal can be upgraded or an article in a "higher-quality" journal can be downgraded.

10   While subject to a degree of subjectivity, these standards for ranking journals and their domain preferences are the same as those used by faculty-tenure review committee in major medical schools in the United States of America in order to evaluate the publication record of a tenure-candidate. Similarly, human experts in various domains recognize that certain information sources can have a predisposition to disregard

15   or highly regard certain authors or types of submitted work. Since the editorial board and peer-reviewers of journals change with time, the tables for grading journals are not static but must be revised over time as reviewers or editors specific to domain specialties change. In combination with the Citation Index of impact journals, these criteria enable the scoring of a reference's support of gene's annotation.

20   Another variable used in the evaluation of the experimental support for an alleged role or function for a gene is a "follow-on" parameter. Reliable experimentalists often will publish a series of papers in reputable journals. They may publish on the same gene or encoded protein ("GeneRef" 230-a attribute of table "FollowOnWork" 230 in Fig. 7, or "ProteinRef" 230-b), a close homolog ("FamilyMemberRef" 230-c), another

25   gene in the same pathway ("PathwayRef" 230-d) or the same gene or pathway in another organism ("altOrganismRef" 230-e). When a large body of work from an individual author or group of authors accumulates, then the probability of "truth" is high. In contrast, a single publication by an author that alleges unusual relationships amongst genes that fails to engender follow-on work (as roughly measured by the Citation Index)

30   by the original author or others has a lower probability of "truth" which is reflected by a lower reference score ("RefScore" 200-k). An intermediate reference score occurs where a single publication triggers much work by other investigators, e.g. a high Citation Index

19

but low "follow-on" value. Thus, this strategy compensates for the overall weakness of
the Citation Index— by merely enumerating the occurrences of a referenced paper, the
Citation Index may not be accurately represent the relatedness of subsequent work.

       Fig. 7 depicts the functional annotative information stored for the genes

5   according to an embodiment of the present invention. Database tables 160, 170, 180,
190, 200, 210, 220, and 230 depicted in Fig. 7 include annotation information derived
from peer-reviewed articles and other information accessed by server 14. A table of the
annotation summary ("AnnotationSummary" 160) includes the sequence name ("SeqFile"
160-a), best hits ("BestHits" 160-b) which refers to the "DNAsequence" table 130

10   ("BestBlastnGID" 130-f), a link to the "Function" table 170 ("Function" 160-c), a link to
the "Role" table 180 ("Role" 160-d), a link to the "Evidence" table 190 ("Evidence" 160-
e). The Function 170, Role 180 and Evidence 190 tables contain many attributes which
all refer to individual References ("Reference" table 200). Any reference in "Reference"
table 200 ("RefID" 200-a) that supports the concept that a gene is an enzyme

15   ("EnzymeRef" 170-a), a receptor ("ReceptorRef" 170-b), a channel or transporter
("ChannelRef" 170-c), a protein interactor ("InteractorRef" 170-d), a structural protein
("StructuralRef" 170-e), a nucleic acid binding protein ("NucleicAcidBindingProtein"
170-f), has a role in cognition ("CognitionRef" 180-a), or a role in development
("DevelopmentRef" 180-b), or a role in endocytosis ("EndocytosisRef" 180-c), a role in

20   exocytosis ("ExocytosisRef" 180-d), or a role in Metabolism ("MetabolismRef" 180-e),
or a role in regulation ("RegulationRef" 180-f), or a role in reproduction
("ReproductionRef" 180-g), or a role in signaling ("SignallingRef" 180-h), or a role in
RNA splicing ("SplicingRef" 180-i), or a role in vesicle trafficking ("TraffickingRef"
180-j), or a role in transcription ("TranscriptionRef" 180-k) is duly linked to the

25   appropriate reference identifier ("RefID" 200-a). The weighted scores for each of these
possible functions is stored as a multi-item list ("FunctionScores" 170-g). Similarly, the
weighted scores for each of the possible roles is stored as a multi-item list; e.g. a
"RoleScores" (180-l) equivalent to "0,100,100,0,0,0,0,0,0,0,0" might correspond to a
single published article on a gene's role in the endocytosis of key nutrients during

30   development in a prominent journal such as Science ("DevelopmentRef" 180-b and
"EndocytosisRef" 180-c). In a database query, such a summary weighted score can be
simply compared to other scores by both the maximum value of each comma-delimited
item as well as the rank order amongst comma-delimited items. Similarly, any

experimental evidence contained in the reference that shows that a gene's encoded protein was immune precipitated ("ImmunePrecipRef" 190-b), a gene's encoded mRNA was hybridized in a Northern assay ("NorthernRef" 190-c), a gene was hybridized in a Southern blot ("SouthernRef" 190-d), a protein band of appropriate predicted size was
5   identified in a Western blot ("WesternRef" 190-e), an open reading frame was identified in a yeast two-hybrid interactor analysis ("InteractorAnalysisRef" 190-f), an enzymatic assay ("BiochemistryRef" 190-g), a pharmacological profile was determined ("PharmacologyRef" 190-h), a predicted homologous domain ("HomologyRef" 190-j) or a predicted structural 3-dimensional motif ("StructureRef" 190-k) is duly referenced to
10   the appropriate reference identifier ("RefID" 200-a).

Referring further to Fig. 7, tables are shown to record the information about any pathway or reference . For any pathway ("Pathway" 210-a in table "Pathway" 210), a role may be assigned ("Role" 210-b), genes of the pathway listed ("GeneList" 210-c) and the location of the pathway identified ("Locations" 210-d). For any reference,
15   a unique identifier ("RefID" 200-a) is recorded, the authors listed ("Author" 200-b), the article title ("Title" 200-c), the journal in which the article was published ("Journal" 200-d), the volume of the journal ("Volume" 200-e), the page numbers of the article ("Page" 200-f), the year of the article's publication ("Year" 200-g), and the reference score link ("RefScore" 200-k). The reference score link 200-k corresponds to the "RefScore"
20   object/table 220 which also contains the reference identifier ("RefID" 220-a), the citation index value (if any) ("CitationIndex" 220-b), the topic field (e.g. immunology or neurobiology) ("Domain" 220-c), a domain weight-adjusted value for the journal quality, as described above, ("JournalRigor" 220-d), and the link to follow-on work table 230 ("FollowOnWork" 220-e). The follow-on table 230 consists of a reference to any
25   subsequent work in which the same gene ("GeneRef" 230-a) or protein ("ProteinRef" 230-b), or homologous gene ("FamilyMemberRef" 230-c), or the same pathway ("PathwayRef" 230-d) or alternate organism ("altOrganismRef" 230-e) was studied by the original investigators.

Referring back to Fig. 3, the present invention then obtains (step 59) and
30   stores (step 60) expression profile data for the genes and their homologs. The expression profile data for a gene describes how the gene is expressed, or transcribed to RNA. Profiles can be created for genes in cells or tissues under influences of a drug, as a cell or tissue develops, or during changes to the physiological state of the cell or tissue, or in response to the development of disease in humans or an animal model. For example, the

21

expression profile data may indicate whether a gene is up-regulated/down-regulated during a stroke.

Fig. 8 depicts the gene expression profile data stored in the database according to an embodiment of the present invention. The four tables depicted in Fig. 8

5   correspond to a summary of the array result conditions ("ArrayResults" 240), the summarized array data ("ArrayData" 250), the details of the probe(s) ("Probe" 260), and the raw data ("RawData" 270). The array result conditions table 240 contains attributes that describe a unique experimental identifier ("ExptID" 240-a), the corresponding bar code ("BarCode" 240-b), the link for probe 1 ("Probe1" 240-c), the link for probe 2

10   ("Probe2" 240-d), a term that describes the grid pattern ("GridPattern" 240-e), the clone set identifier ("CloneSet" 240-f), the link to array data ("ArrayData" 240-g), and a comment ("Comment" 240-h). The array data table 250 contains attributes to describe the experimental identifier ("ExptID" 250-a), the name of the cDNA sequence ("seqFile" 250-b), the arithmetic mean of the background or normalized data ("Mean" 250-c), the

15   standard deviation ("StdDev" 250-d), the ratio of any paired means derived from simultaneous application of two probes ("Ratio" 250-e), the time point at which the probes were made ("TimePt" 250-g), the biological state (e.g. diseased or normal) of the probe's mRNA origin ("State" 250-h), the clustering method ("ClusterMethod" 250-i), the cluster number ("Cluster" 250-j), the total number of clusters ("TotalClusters" 250-k),

20   the cluster order pattern derived from the auto-regression analysis used in the causality analysis ("ClusterOrder" 250-l) and the date of the clustering ("ClusterDate" 250-m).

The probe data table 260 contains attributes for the probe identifier ("ProbeID" 260-a), the date of probe generation ("Date" 260-b), the type (first strand cDNA or double-stranded cDNA) of probe ("Type" 260-c), the biological model

25   ("Model" 260-d), the identifier for the preparation of RNA ("RNAprep" 26-e), the labeling (radioactive or fluorescent) method ("LabelType" 260-f), the time point at which the RNA was collected ("TimePt" 250-g), the biological state of the probe's mRNA origin ("State" 250-h), and a comment ("Comment" 260-i).

The raw data table 270 contains attributes for the experimental identifier

30   ("ExptID" 270-a), the sequence name ("seqFile" 270-b), the probe name ("Probe" 270-c), the raw intensity value ("RawValue" 270-d), the local background or normalization factor ("LocalBgnd/factor" 270-e), and the arithmetically corrected intensity value ("CorrectedValue" 270-f).

22

Referring back to Fig. 3, the present invention then performs clustering

analysis on the behavior of DNA sequences in expression profile studies (step 62).

According to clustering analysis, data complexity is reduced by partitioning the genes

into groups or "clusters" that have similar attributes. These attributes can be the behavior

5   of genes monitored over multiple time points in response to an injury, onset of disease or

altered physiological state (e.g. intensity or ratio of intensities resulting from

hybridization of a gene set with probes derived from normal and diseased tissue). Also,

these attributes can simply be the response of genes from cells, tissues or animals treated

with multiple concentrations (e.g. 5, 6 or 7 concentrations) of many drugs (e.g. 10, 100,

10   1000 or 10,000) with differing mechanisms of action at a single time point. These

attributes can also be the response of cells or animals subjected to many altered

physiological states (e.g. elevated or diminished nutrients, ions or temperature, transient

ischemia, shock, anxiety, discomfort or depression) monitored at a single time point

relative to untreated cells or tissues. The result of clustering gene expression data are

15   clusters of genes with similar expression profiles.

An embodiment of the present invention implements a method of gene

clustering that is tuned to the simplified, yet specific nature of the array data itself. In

order to reduce data complexity, many clustering methods have been applied to gene

expression profile data: these include hierarchical, K-means, self-organizing maps

20   (Tamayo et al. PNAS 96:2907-12), or support vector machines (M. Brown et al. PNAS

97:262-7). An embodiment of the present invention uses a K-means distance with

Euclidean distance or other distance metrics (provided by Partek of St. Louis MO)

because of its ability to efficiently cluster data in an automated unsupervised manner.

One of the common criticisms of K-means clustering is that the number of clusters must

25   be determined a priori. However, the present invention uses the Davies-Bouldin

algorithm (IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAM1-

1, April 1979) which determines the optimal number of clusters based upon the dispersion

and flatness of clusters.

According to an embodiment of the present invention, the present

30   invention may cluster the genes based on time-course data as described by the expression

profile data. According to a specific embodiment of the present invention, packages

provided by Partek Inc. and/or SAS Institute, Incorporated of Cary, North Carolina may

be used to perform the clustering analysis. For time-course data, the clustering analysis

23

may also include causality analysis to predict ordered relationships between clusters on a time basis. Causality analysis is performed using a regressive method performed with software packages such as the Statistical Analysis Software from SAS Institute, Incorporated. The results from the clustering analysis are stored in a database (step 64).

5   The cluster analysis results are inserted into the array data table 250 of Fig. 8: for each gene ("seqFile" 250-b), the clustering method ("ClusterMethod" 250-i), a cluster number ("Cluster" 250-j), the total number of clusters ("TotalClusters" 250-k), and the cluster order ("ClusterOrder" 250-l).

       The type of clustering method(s) used to analyze array data depends upon
10   (a) a priori knowledge about the behavior of the immobilized genes, (b) the composition of the gene set itself, and (c) the choice of array technologies. Array technologies come in two general forms: cDNA and oligonucleotide arrays. Since the Affymetrix arrays often have a higher density than cDNA arrays, the emphasis has been to increase the number of sequences per unit surface area in order to gain thoroughness. Often times,
15   inadequate attention is paid to the design of the actual DNA attached to the array. Thus, many array chip designs seek to deposit large numbers of gene fragments per chip; such as species-specific chips (mouse, rat or human chips from Affymetrix, Santa Clara, CA) or genes representative of a field (apoptosis, cancer or neurobiology chips from Affymetrix or Clonetech, Palo Alto, CA. However, analysis of such chips is complicated
20   by the fact that most genes on the chip may be irrelevant to the biological system being studied.

       According to the present invention, the analysis of gene clusters is vastly simplified by the immobilization of a plurality of genes that are actually disease- or physiologically- specific. Such collections of genes can be generated by any method that
25   enables the identification of genes expressed at a measurable level higher in one state than another. For example, in tumors or animals subjected to ischemia, those skilled in the art of molecular cloning can identify and isolate cDNA clones and derive the sequences thereof for genes whose expression is elevated 2, 3 or 10 fold higher in the altered physiological state; e.g. differential display and subtractive cloning are two such methods.
30   The number of disease-related or physiologically-related genes may range from 1000, 6000, 10,000, or 20,000 per chip.

24

When analyzed by principal components analysis, typically 90% of the variability in the gene expression profile data generated by arrays of 6000-10,000 disease- or paradigm- specific cDNA targets can be explained by the first 3 principal components or eigenvectors. With a large number of genes unrelated to the biological paradigm of the
5   probe (e.g. 40,000- 60,000 genes present on some Affymetrix arrays), the data variability is likely explained by many more principal components which makes it difficult to analyze more than any 3 of all principal components in 3-dimensional space. For these instances, other clustering methods might be more appropriate, such as hierarchical clustering. However, optimal hierarchical clustering is highly iterative and false clusters
10   are often generated.

In order to infer the time-order of gene clusters derived from the above, it is possible to calculate likely causality by a moving auto-regressive analysis. A time-order is a linear ranking of clusters by a deduced set of relationships ordering the first possible cluster relative to other clusters in an iterative process. A biological example of
15   this problem is the goal of understanding which genes respond earliest to an injury or infection followed by the elucidation of time of activation of subsequent, related or unrelated genes. A ordered set of clusters from expression profile data is achieved initially by selecting a representative subset of genes near the centroid of each cluster (e.g. 2, 5 or 10 representing 1-10% of the total number of genes) and performing a
20   moving auto-regressive test against the remaining genes of the monitored population of genes (e.g. 2, 5 or 10 genes compared to all 6000 or 10,000 genes) from all clusters (Statistical Analysis Software of SAS Institute, Incorporated, Cary, North Carolina). The ranked order of clusters is stored in "ClusterOrder" (250-1) in step 64.

The accuracy of ordering clusters is dependent on the completeness of the
25   calculation, but calculation of cluster order is computationally intensive. For example. according to a specific embodiment, the above calculation requires about 24 hours on a standard single CPU Unix workstation with 1 gigabyte of RAM; e.g. a Sun Ultra10 workstation with 300 MHz CPU. This time-series analysis is only applicable to datasets with regularly spaced time-points (e.g. 10, 20 or 40 instances spaced 30 min, 1 hr or 3 hrs
30   apart). The time-resolution of the causality analysis is dependent upon the density of intervals over the entire course experimental course. For the highest resolution of time-ordered relationships amongst clusters, 20, 50, or 100 time-points are preferable. For the

25

highest accuracy amongst clusters, a comprehensive auto-regression is calculated
provided sufficient computer power (e.g. 6000 genes compared to 6000 genes or 10,000
genes compared to 10,000 genes requires supercomputer ability or the efforts of a cluster
of workstations such as Beowulf: (http://www.beowulf.org/)).

5   Referring back to Fig. 3, after the clustering analysis, the present invention
may obtain pathway information (step 65) for the genes and their homologs and store the
pathway information to the database (step 66). Pathway information can be accessed
from public pathway databases such as the Kyoto Encyclopedia of Genes and Genomes
(KEGG) or the Munich Information Center for Protein Sequences (MIPS), or derived

10   from the literature using information extraction methods, as described earlier.

               According to an embodiment of the present invention, the database used
for storing information associated with the genes correlates the annotative information
with numerical gene expression profile data. Within each time-resolved cluster of genes
with similar behavior, multiple types of genes may exist ("Cluster" 250-j is linked to

15   "seqFile" 250-b which can be referenced to the annotation summary "seqFile" 160-a).
For example, genes that are stimulated immediately after an injury or stress might include
chaperones or heat shock proteins in order to prevent misfolded proteins. Similarly,
transcription factors might be triggered to increase the production of protective systems.
All of these genes' mRNA levels could be elevated within the first 5 minutes post-injury

20   but their mRNA levels might diminish at varying rates. Subsequently, secondary and
tertiary groups of genes might be activated in response to the transcription factors. While
the clustering and causality analysis described above can identify groups of early onset
genes, it cannot distinguish the functional relationship, if any, between differing kinds of
genes within each time-ordered group. For this task, integration of the annotation of all

25   genes for each time-ordered group is necessary. Currently, such analyses are performed
by human experts and are limited by recall while a database query constrained by user-
defined parameters could present all possible cross-connections that are likely or less
likely– depending upon the reliability threshold ("FunctionScores" 170-g, "RoleScores"
180-l) for "truth" of a relationship defined by the user. Thus, multiple alternate scenarios

30   can be presented in a database or in tabular form or graphical objects linked by lines that
purport directional control and annotative text describing the likelihood of the interaction

26

along with hyperlinks to relevant published articles via HTML (hypertext markup language) methods.

A feature of the present invention is that it provides support for both intra- and inter- time-resolved gene cluster components; i.e. between or amongst genes in
5   subsequent or previous groups of genes. Thus, a human expert can choose from a palette of options to refine a first iteration of gene network or pathway building. The parameters in turn can be used to recalculate the likelihood of other annotations and pathways to explain the behavior of a single gene, group of genes, or cluster of genes. Collectively, these methods can reduce the number of differentially regulated genes to a smaller
10   plurality; from which candidate genes can be chosen by the human expert.

The information stored in the database according to the present invention facilitates the identification of candidate genes (step 68 in Fig. 3). Identification of candidate genes results from the merge of the time-ordered gene expression clusters and the function(s), role(s) and/or pathway(s) information of the cluster members. The
15   reference score-based assignments for either majority or minority view annotations of function(s), role(s) and/or pathway(s) enables the identification of new or serendipitous relationships. Such biological novelty, i.e. the unexpected up- or down-regulation of a gene in the context of an existing or new pathway, is one of the hallmarks of candidate genes. For example, in a signaling pathway, study of a disease model may reveal that
20   one, two or three known phosphodiesterases are up-regulated in the context of a pathway not normally characterized by those enzymes. Or, a new family member of this enzyme class might be discovered up-regulated along with the expected enzyme. Both are examples of candidate genes revealed by the combination of annotated DNA sequences and expression profiling data– particularly if the published literature contained an obscure
25   reference to such a relationship under abnormal circumstances dissimilar to the conditions of the experimental paradigm. The latter result would be significant due to the redundancy of biological systems. Conversely, if 7, 8 or 9 of 10 genes of a well known pathway are found to be up-regulated in a disease or injury model (as determined by a comparison of all pathways of each gene expression profile cluster), then the 1, 2 or 3
30   genes that failed to be induced (as determined by a query comparison to the pathway database) might also be considered candidate genes. In this example, the user might conclude that a new inhibitor is blocking the 1, 2, or 3 missing genes and hence blocking

27

the inhibitor might diminish the pathology or improve recovery. The user might then search for known or postulated inhibitors of any member of the pathway.

The information stored in the database may be accessed or queried by users interested in identifying candidate genes. According to a specific embodiment, the

5   present invention provides an interface allowing users to specify a query including criteria characterizing candidate genes. In response to the user query, the present invention searches the database to identify genes which satisfy the user-specified search criteria. A typical search might examine the group of classified genes (e.g. by function, role or pathway) appearing in an early or middle expression cluster (based on "Cluster" 250-j

10   and "ClusterOrder" 250-l). By comparing the similar attributes (e.g. a query of the type "what apoptotic regulator genes are present in early clusters along chemokine genes?") within upstream or downstream clusters, the user may be able to deduce, for example, that the apoptotic pathway in a particular infection model of immune cells was altered by either (a) the appearance of a new apoptotic regulator gene or chemokine at an

15   unexpected time or cluster, or (b) the absence of altered expression a gene known to be induced in the pathway. Alternatively, the user might query what low-likelihood roles or pathways might explain the presence of a given class of receptors. In response to the user query, the present invention uses the user-specified query criteria to search the information stored in the database and outputs genes which satisfy the user-specified

20   search criteria by either their presence or omission from either known or low- likelihood roles (or pathways) or lists of genes with known function(s) or role(s). In this manner, the information stored for the plurality of DNA sequences and their behavior in expression profile data facilitates identification of candidate genes.

Although specific embodiments of the invention have been described,

25   various modifications, alterations, alternative constructions, and equivalents are also encompassed within the scope of this application. The described invention is not restricted to operation within certain specific data processing environments, but is free to operate within a plurality of data processing environments. For example, although the present invention has been described in a distributed computer network environment, the

30   present invention may also be incorporated in a single stand-alone computer system. In such an environment, the same stand-alone computer has access to the various biological databases according to the present invention and may act both as a client and a server.

28

Additionally, although the present invention has been described using a particular series of transactions and steps, it should be apparent to those skilled in the art that the scope of the present invention is not limited to the described series of transactions and steps.

Further, while the present invention has been described using a particular

5   combination of hardware and software, it should be recognized that other combinations of hardware and software are also within the scope of the present invention. The present invention may be implemented only in hardware or only in software or using combinations thereof.

The specification and drawings are, accordingly, to be regarded in an

10   illustrative rather than a restrictive sense. It will, however, be evident that additions, subtractions, deletions, and other modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the claims.

WHAT IS CLAIMED IS:

1   1.   A computer-implemented method of identifying candidate genes
2   from a plurality of DNA sequences, the method comprising:
3   obtaining results of a homology search for the plurality of DNA sequences,
4   the homology search results comprising information about homologs of the plurality of
5   DNA sequences;
6   obtaining annotative information for the plurality of DNA sequences, the
7   annotative information comprising information about the biochemical functions and
8   physiological roles of the plurality of DNA sequences;
9   obtaining gene expression profile data for the plurality of DNA sequences,
10   the gene expression profile data describing behavioral patterns of the plurality of DNA
11   sequences;
12   clustering the plurality of DNA sequences based on the behavioral patterns
13   of the plurality of DNA sequences as described by the gene expression profile data;
14   storing the results of the homology search, the annotative information, the
15   gene expression profile data, and results from clustering the plurality of DNA sequences
16   in a database;
17   receiving a query identifying criteria for the candidate genes; and
18   searching the database, in response to the query, to identify a set of DNA
19   sequences from the plurality of DNA sequences which satisfy the query criteria.

1   2.   The method of claim 1 wherein the homology search for the
2   plurality of DNA sequences comprises performing BLAST analysis, Smith-Waterman
3   analysis, Hidden Markov Model (HMM) analysis, and EMotif analysis.

1   3.   The method of claim 2 wherein performing the BLAST analysis,
2   the Smith-Waterman analysis, the Hidden Markov Model (HMM) analysis, and the
3   EMotif analysis comprises:
4   performing the BLAST analysis on the first plurality of DNA sequences
5   using a first database of sequences;
6   identifying a second plurality of DNA sequences from the first plurality of
7   sequences which are not known based on the BLAST analysis using the first database of
8   sequences;

9   performing Smith-Waterman analysis on the second plurality of DNA
10   sequences using a protein database and a translated patent database;
11   identifying a third plurality of DNA sequences from the second plurality of
12   sequences which are not known based on the Smith-Waterman analysis;
13   performing Hidden Markov Model (HMM) analysis and EMotif analysis
14   on the third plurality of DNA sequences using the protein database and GenBank
15   database; and
16   performing BLAST analysis on the third plurality of DNA sequences using
17   GenBank EST database.

1   4.   The method of claim 1 wherein obtaining the annotative
2   information comprises:
3   identifying known genes from the first plurality of DNA sequences based
4   on the homology search; and
5   accessing information sources storing annotative information for the
6   known genes; and
7   extracting the annotative information from the information sources for the
8   known genes.

1   5.   The method of claim 4 wherein extracting the annotative
2   information comprises:
3   assigning a reference score to the extracted annotative information based
4   on the level of acceptance of the role or function of the known genes as described by the
5   annotative information such that annotative information with a high level of acceptance is
6   assigned a higher reference score than annotative information with a low level of
7   acceptance.

1   6.   The method of claim 4 wherein the information sources include
2   GenBank database, SWISS-PROT database, Medline database, and biomedical
3   publications.

1   7.   The method of claim 4 wherein:
2   accessing the information sources comprises accessing biomedical
3   publications;
4   extracting the annotative information comprises:

31

5   for annotative information extracted from each biomedical

6   publication:

7   assigning a reference score to the extracted annotative

8   information based on characteristics of the biomedical publication, the reference score

9   indicating the level of acceptance of the role or function of the known genes as described

10 by the annotative information extracted from the biomedical publication; and

11   storing the annotative information in the database comprises storing the

12 reference score.

1   8.   The method of claim 7 wherein assigning the reference score

2   comprises:

3   using a score derived from a citation index database to calculate the

4   reference score, the score derived from the citation index database indicating the number

5   of times that the annotative information from the biomedical publication was referenced

6   by other information sources.

1   9.   The method of claim 7 wherein assigning the reference score

2   further comprises:

3   ranking the biomedical publications; and

4   assigning the reference score to the annotative information extracted from

5   the biomedical publication based on the ranking of the biomedical publication.

1   10.   The method of claim 1 wherein clustering the plurality of DNA

2   sequences comprises determining relationships between clusters of DNA sequences from

3   the plurality of DNA sequences.

1   11.   The method of claim 1 wherein clustering the plurality of DNA

2   sequences comprises clustering the plurality of DNA sequences based on time-course

3   data described by the gene expression profile data.

1   12.   The method of claim 1 wherein storing the information in the

2   database comprises correlating the annotative information for the plurality of DNA

3   sequences with the gene expression profile data for the plurality of DNA sequences.

1   13.   A method of identifying candidate genes comprising:

2   configuring a query identifying criteria for the candidate genes;

32

3   communicating the query to a server storing information related to a

4   plurality of DNA sequences, the information comprising:

5   results of a homology search for the plurality of DNA sequences,

6   the homology search results comprising information about homologs of the plurality of

7   DNA sequences;

8   information about the biochemical functions and physiological

9   roles of the plurality of DNA sequences;

10   information describing behavioral patterns of the plurality of DNA

11   sequences; and

12   results from clustering the plurality of DNA sequences based on

13   the behavioral patterns of the plurality of DNA sequences as described by the gene

14   expression profile data; and

15   receiving from the server, in response to the query, a first set of DNA

16   sequences from the plurality of DNA sequences, wherein the first set of DNA sequences

17   satisfy the criteria for the candidate genes identified in the query.

1   14.   A data processing system for identifying candidate genes from a

2   plurality of DNA sequences, the system comprising:

3   a processor; and

4   a memory coupled to the processor, the memory configured to store

5   instructions for execution by the processor, the instructions comprising:

6   instructions for obtaining results of a homology search for the

7   plurality of DNA sequences, the homology search results comprising information about

8   homologs of the plurality of DNA sequences;

9   instructions for obtaining annotative information for the plurality of

10   DNA sequences, the annotative information comprising information about the

11   biochemical functions and physiological roles of the plurality of DNA sequences;

12   instructions for obtaining gene expression profile data for the

13   plurality of DNA sequences, the gene expression profile data describing behavioral

14   patterns of the plurality of DNA sequences;

15   instructions for clustering the plurality of DNA sequences based on

16   the behavioral patterns of the plurality of DNA sequences as described by the gene

17   expression profile data;

33

18   instructions for storing the results of the homology search, the

19   annotative information, the gene expression profile data, and results from clustering the

20   plurality of DNA sequences in the memory; and

21   instructions for searching the information stored in the memory, in

22   response to a query identifying criteria for the candidate genes, to identify a set of DNA

23   sequences from the plurality of DNA sequences which satisfy the query criteria.

1   15.   The system of claim 14 wherein the memory is further configured

2   to store instructions for performing the homology search, the instructions comprising:

3   instructions for performing BLAST analysis on the first plurality of DNA

4   sequences using a first database of sequences;

5   instructions for identifying a second plurality of DNA sequences from the

6   first plurality of sequences which are not known based on the BLAST analysis using the

7   first database of sequences;

8   instructions for performing Smith-Waterman analysis on the second

9   plurality of DNA sequences using a protein database and a translated patent database;

10   instructions for identifying a third plurality of DNA sequences from the

11   second plurality of sequences which are not known based on the Smith-Waterman

12   analysis;

13   instructions for performing Hidden Markov Model (HMM) analysis and

14   EMotif analysis on the third plurality of DNA sequences using the protein database and

15   GenBank database; and

16   instructions for performing BLAST analysis on the third plurality of DNA

17   sequences using GenBank EST database.

1   16.   The system of claim 14 wherein the instructions for obtaining the

2   annotative information comprise:

3   instructions for identifying known genes from the first plurality of DNA

4   sequences based on the homology search; and

5   instructions for accessing information sources storing annotative

6   information for the known genes; and

7   instructions for extracting the annotative information from the information

8   sources for the known genes.

34

1   17.   The system of claim 16 wherein the instructions for extracting the
2   annotative information comprise:
3   instructions for assigning a reference score to the extracted annotative
4   information based on the level of acceptance of the role or function of the known genes as
5   described by the annotative information such that annotative information with a high level
6   of acceptance is assigned a higher reference score than annotative information with a low
7   level of acceptance.

1   18.   The system of claim 16 wherein the information sources include
2   GenBank database, SWISS-PROT database, Medline database, and biomedical
3   publications.

1   19.   The system of claim 16 wherein:
2   the instructions for accessing the information sources comprise
3   instructions for accessing biomedical publications;
4   the instructions for extracting the annotative information comprise:
5   instructions for assigning a reference score to annotative
6   information extracted from each biomedical publication based on characteristics of the
7   biomedical publication, the reference score indicating the level of acceptance of the role   .
8   or function of the known genes as described by the annotative information extracted from
9   the biomedical publication; and
10   the instructions for storing the annotative information in the memory
11   comprise instructions for storing the reference score.

1   20.   The system of claim 19 wherein the instructions for assigning the
2   reference score comprise:
3   instructions for using a score derived from a citation index database to
4   calculate the reference score, the score derived from the citation index database indicating
5   the number of times that the annotative information from the biomedical publication was
6   referenced by other information sources.

1   21.   The system of claim 19 wherein the instructions for assigning the
2   reference score comprise:
3   instructions for ranking the biomedical publications; and

4   instructions for assigning the reference score to the annotative information

5   extracted from the biomedical publication based on the ranking of the biomedical

6   publication.

1   22.   The system of claim 14 wherein the instructions for clustering the

2   plurality of DNA sequences comprise instructions for determining relationships between

3   clusters of DNA sequences from the plurality of DNA sequences.

1   23.   The system of claim 14 wherein the instructions for clustering the

2   plurality of DNA sequences comprise instructions for clustering the plurality of DNA

3   sequences based on time-course data described by the gene expression profile data.

1   24.   The system of claim 14 wherein the instructions for storing the

2   information in the database comprise instructions for correlating the annotative

3   information for the plurality of DNA sequences with the gene expression profile data for

4   the plurality of DNA sequences.

1   25.   A system for identifying candidate genes comprising:

2   a communication network;

3   a first computer coupled to the communication network; and

4   a second computer coupled to the communication network, the second

5   computer configured to store:

6   results of a homology search for a plurality of DNA sequences, the

7   homology search results comprising information about homologs of the plurality of DNA

8   sequences;

9   information about the biochemical functions and physiological

10   roles of the plurality of DNA sequences;

11   information describing behavioral patterns of the plurality of DNA

12   sequences; and

13   results from clustering the plurality of DNA sequences based on

14   the behavioral patterns of the plurality of DNA sequences as described by the gene

15   expression profile data;

16   wherein the first computer is configured to communicate a query to the

17   second computer, the query identifying criteria for the candidate genes; and

36

18   wherein the first computer is configured to receive from the second

19   computer, in response to the query, a first set of DNA sequences from the plurality of

20   DNA sequences which satisfy the criteria for the candidate genes identified in the query.

1   26.   A computer program product stored on a computer-readable

2   storage medium for identifying candidate genes from a plurality of DNA sequences, the

3   computer program product comprising:

4   code for obtaining results of a homology search for the plurality of DNA

5   sequences, the homology search results comprising information about homologs of the

6   plurality of DNA sequences;

7   code for obtaining annotative information for the plurality of DNA

8   sequences, the annotative information comprising information about the biochemical

9   functions and physiological roles of the plurality of DNA sequences;

10   code for obtaining gene expression profile data for the plurality of DNA

11   sequences, the gene expression profile data describing behavioral patterns of the plurality

12   of DNA sequences;

13   code for clustering the plurality of DNA sequences based on the

14   behavioral patterns of the plurality of DNA sequences as described by the gene

15   expression profile data;

16   code for storing the results of the homology search, the annotative

17   information, the gene expression profile data, and results from clustering the plurality of

18   DNA sequences in a database;

19   code for receiving a query identifying criteria for the candidate genes;

20   code for searching the database, in response to the query, to identify a set

21   of DNA sequences from the plurality of DNA sequences which satisfy the query criteria.

1   27.   A computer program product stored on a computer-readable

2   storage medium for identifying candidate genes, the computer program product

3   comprising:

4   code for configuring a query identifying criteria for the candidate genes;

5   code for communicating the query to a server storing information related

6   to a plurality of DNA sequences, the information comprising:

37

7   results of a homology search for the plurality of DNA sequences,

8   the homology search results comprising information about homologs of the plurality of

9   DNA sequences;

10   information about the biochemical functions and physiological

11   roles of the plurality of DNA sequences;

12   information describing behavioral patterns of the plurality of DNA

13   sequences; and

14   results from clustering the plurality of DNA sequences based on

15   the behavioral patterns of the plurality of DNA sequences as described by the gene

16   expression profile data; and

17   code for receiving from the server, in response to the query, a first set of

18   DNA sequences from the plurality of DNA sequences, wherein the first set of DNA

19   sequences satisfy the criteria for the candidate genes identified in the query.

Fig. 1

Fig. 2

50

Fig. 3

Fig. 4

Fig. 5

140

142

Access information sources containing functional information for the sequences

144

Access databases

Extract functional information

146

Summarize the extracted functional information

148

Store the functional information in the database

150

Fig. 6

(A) AnnotationSummary (160)
- seqFile – 160-a
- BestHits – 160-b
- Function – 160-c
- Role – 160-d
- Evidence – 160-e
- Locations – 160-f
- Pathway – 160-g
- Comment – 160-h
- AnnotatorName – 160-i
- AnnotationDate – 160-j

(F) Pathway (210)
- Pathway – 210-a
- Role – 210-b
- GeneList – 210-c
- Locations – 210-d

(B) Function (170)
- EnzymeRef – 170-a
- ReceptorRef – 170-b
- ChannelRef – 170-c
- InteractorRef – 170-d
- StructuralRef – 170-e
- NucleicAcidBindingProteinRef (170-f)
- FunctionScores – 170-g

(E) Reference (200)
- RefID – 200-a
- Author – 200-b
- Title – 200-c
- Journal – 200-d
- Volume – 200-e
- Page – 200-f
- Year – 200-g
- Comment – 200-h
- AnnotatorName – 200-i
- AnnotationDate – 200-j
- RefScore – 200-k

(C) Role (180)
- CognitionRef – 180-a
- DevelopmentRef – 180-b
- EndocytosisRef – 180-c
- ExocytosisRef – 180-d
- MetabolismRef – 180-e
- RegulationRef – 180-f
- ReproductionRef – 180-g
- SignallingRef – 180-h
- SplicingRef – 180-i
- TraffickingRef – 180-j
- TranscriptionRef – 180-k
- RoleScores – 180-L

(G) RefScore (220)
- RefID – 220-a
- CitationIndex – 220-b
- Domain – 220-c
- JournalRigor – 220-d
- FollowOnWork – 220-e
- Comment – 220-f

(D) Evidence (190)
- Experimental – 190-a
  - ImmunePrecipRef – 190-b
  - NorthernRef – 190-c
  - SouthernRef – 190-d
  - WesternRef – 190-e
  - InteractorAnalysisRef – 190-f
  - BiochemistryRef – 190-g
  - PharmacologyRef – 190-h
- Predicted – 190-i
  - HomologyRef – 190-j
  - StructureRef – 190-k
- Comment – 190-L
- Date – 190-m

(H) FollowOnWork (230)
- GeneRef – 230-a
- ProteinRef – 230-b
- FamilyMemberRef – 230-c
- PathwayRef – 230-d
- altOrganismRef – 230-e
- Comment – 230-f

Fig. 7

(A) ArrayResults (240)

- 240-a — ExptID
- 240-b — BarCode
- Probe1 ~ 240-c
- Probe2 ~ 240-d
- 240-e — GridPattern
- 240-f — CloneSet
- 240-g — ArrayData
- 240-h — Comment

(B) ArrayData (250)

- 250-a — ExptID
- 250-b — seqFile
- 250-c — Mean
- 250-d — StdDev
- 250-e — Ratio
- 250-f — RawData
- 250-g — TimePt
- 250-h — State
- 250-i — ClusterMethod
- 250-j — Cluster
- 250-k — TotalClusters
- 250-l — ClusterOrder
- 250-m — ClusterDate

(C) Probe (260)

- 260-a — ProbeID
- 260-b — Date
- 260-c — Type
- 260-d — Model
- 260-e — RNAprep
- 260-f — LabelType
- 260-g — TimePt
- 260-h — State
- 260-i — Comment

(D) RawData (270)

- 270-a — ExptID
- 270-b — seqFile
- 270-c — Probe
- 270-d — RawValue
- 270-e — LocalBgnd/factor
- 270-f — CorrectedValue

Fig. 8

Score	Journal	Score	Journal
100	Nature	80	Brain
100	Science	80	Journal of Physiology
97	Cell	80	Molecular and Cellular Endocrinology
96	Neuron	80	Molecular and Cellular Neurosciences
96	Nature Structural Biology	78	Genome Research
95	Nature Cell Biology	78	Journal of Clinical Epidemiology
95	Nature Genetics	78	Pharmacology
95	Nature Medicine	75	Genomics
95	Nature Neuroscience	75	Journal of Cellular Biochemistry
94	Journal of Structural Biology	75	Journal of Investigative Dermatology
94	New England Journal of Medicine	75	Journal of Medical Genetics
92	Journal of Biological Chemistry	75	Journal of Medical Virology
92	Journal of Lipid Research	75	Journal of Psychiatric Research
92	Journal of Molecular Biology	75	Nitric Oxide
92	Neuroscience	75	Toxicology
91	Journal of Bacteriology	74	Brain and Development
91	Journal of Cell Biology	73	Cancer Immunology, Immunotherapy
91	Journal of Immunology	70	Brain Injury
91	Journal of Neurochemistry	70	Cytogenetics and Cell Genetics
90	Biochemistry	70	Glia
90	Bioinformatics	70	Journal of Biotechnology
90	Immunology	70	Obesity Research
90	Journal of Clinical Investigation	69	Immunopharmacology
90	Journal of Internal Medicine	68	Cellular Signalling
90	Journal of Medicinal Chemistry	68	Immunology and Cell Biology
90	Journal of Membrane Biology	68	Stem Cells
90	Journal of Neuroscience	68	Stroke
90	Journal of Virology	68	Synapse
90	Molecular Microbiology	67	Immunology Letters
89	Cell Biology	65	Cell Motility and the Cytoskeleton
89	Genetics	65	Gene
89	Journal of Experimental Medicine	65	Hormone Research
89	Journal of Neurobiology	65	Journal of Cognitive Neuroscience
89	Neurology	65	Journal of the Neurological Sciences
88	Infection and Immunity	65	Pharmacological Research
87	Journal of Neuroendocrinology	63	American Journal of Physiology
87	Journal of Neuroimmunology	62	American Journal of Medicine
87	Journal of Neurology	62	Journal of Comparative Neurology
87	Neuroendocrinology	60	American Journal of Human Genetics
86	Neuroscience Letters	60	Cell Proliferation
85	Development	55	Journal of Affective Disorders
85	Proceedings National Academy of Sciences	50	DNA and Cell Biology
85	Immunogenetics	45	Anesthesiology
85	Journal of Cell Science	41	Biochemical Journal
85	Journal of Infectious Diseases	40	Anesthesia and Analgesia
85	Journal of Neurophysiology	40	Biochemical Pharmacology
85	Molecular Pharmacology	40	Cell Biochemistry and Function
85	Molecular and Cellular Biology	40	Cell and Tissue Research
84	Brain Research	40	Cerebral Cortex
84	Developmental Biology	40	Experimental Cell Research
84	Developmental Neuroscience	40	Histochemistry and Cell Biology
84	Molecular Biology of the Cell	35	Journal of Dermatological Science
83	Cancer	20	Chemistry and Biology
82	Analytical Biochemistry	20	Genes, Chromosomes and Cancer
82	Journal of Endocrinology	20	International Journal of Biochemistry and Cell Biology
82	Journal of Experimental Biology	20	International Journal of Developmental Neuroscience
82	Journal of Histochemistry and Cytochemistr	10	Archives of Biochemistry and Biophysics
82	Journal of Molecular Endocrinology	10	International Journal of Antimicrobial Agents
82	Molecular Immunology	10	International Journal of Cancer
82	Natural Immunity	10	International Journal of Experimental Pathology
82	Nucleic Acids Research	10	Journal of Antisense Research

Fig. 9

INTERNATIONAL SEARCH REPORT

A.   CLASSIFICATION OF SUBJECT MATTER
IPC(7)   :   G01N 31/00; G06F 17/00, 15/00
US CL   :   702/27; 706/47; 712/200
According to International Patent Classification (IPC) or to both national classification and IPC

B.   FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
    U.S. : 702/27; 706/47; 712/200

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C.   DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	MARCOTTE et al. A combined algorithm for genomewide prediction of protein function. Nature. 04 November 1999, Vol. 402, pages 83-86, see entire document	1-13
Y,P	ENRIGHT et al. Protein interaction maps for complete genomes based on gene fusion events. Nature. 04 November 1999, Vol. 402, pages 86-90, see entire document	1-13
A	WO 99/05591 A2 (AFFYMETRIX, INC.) 04 February 1999 (04-02-99), Abstract, pages 3-4.	14-24
A	WO 99/05574 A1 (AFFYMETRIX, INC.) 04 February 1999 (04-02-99), Abstract, Figures 2A, 2B, pages 10-15.	25-27

☐ Further documents are listed in the continuation of Box C.   ☐   See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier application or patent published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 September 2000 (22.09.2000)	**09 JAN 2001**

Name and mailing address of the ISA/US
        Commissioner of Patents and Trademarks
        Box PCT
        Washington, D.C. 20231
Facsimile No. (703)305-3230
	Authorized officer **TERRY J. DEY**
	Young J. Kim **PARALEGAL SPECIALIST**
	          TECHNOLOGY CENTER 1600
	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)

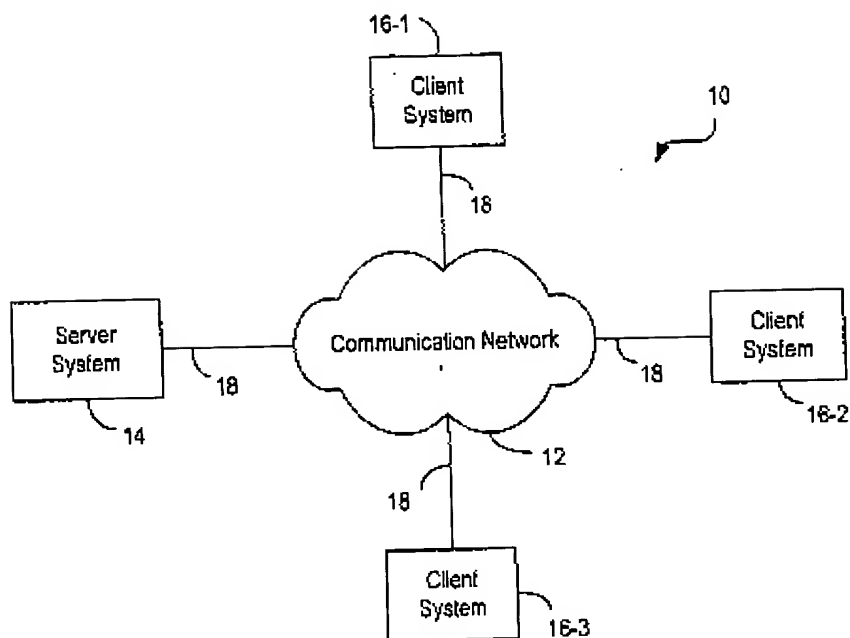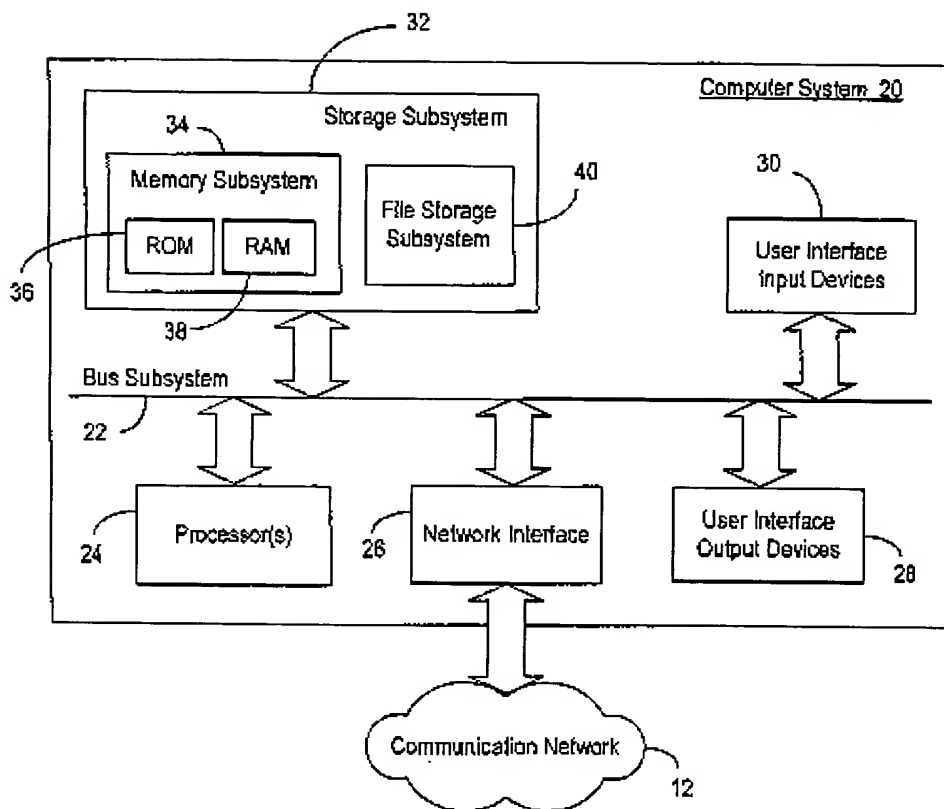**Continuation of B. FIELDS SEARCHED Item 3:** STN Commercial Database (Biosis, Medline, Embase, Embal, SciSearch, BiotechDS, CAPlus)

Search terms: RDBMS, Relational Database, Database, BLAST, HMM, Smith Waterman, clustering, EST, expression profile

Fig. 1

Fig. 2

50

Access homology search results for a plurality of DNA sequences
52

Extract information from the homology search results, organize, and store the data in a particular format (e.g. in a database)
54

Obtain descriptive annotative information related to gene functions and roles
56

Store the functional annotations in the database
58

Obtain expression profile data for the genes
59

Store the expression profile data in the database
60

Perform clustering analysis
62

Store clustering analysis results in the database
64

Obtain pathway information
65

Store the pathway information in the database
66

Facilitate identification of candidate genes
68

Fig. 3

Fig. 4

Fig. 5

140

142
Access information sources
containing functional
information for the sequences

144
Access databases

Extract functional information
146

Summarize the extracted functional information
148

Store the functional information in the database
150

Fig. 6

Fig. 7

(A) ArrayResults (240)
240-a ExptID
240-b BarCode
Probe1 ~ 240-c
Probe2 ~ 240-d
240-e GridPattern
240-f CloneSet
240-g ArrayData
240-h Comment

(B) ArrayData (250)
250-a ExptID
250-b seqFile
250-c Mean
250-d StdDev
250-e Ratio
250-f RawData
250-g TimePt
250-h State
250-i ClusterMethod
250-j Cluster
250-k TotalClusters
250-l ClusterOrder
250-m ClusterDate

(C) Probe (260)
260-a ProbeID
260-b Date
260-c Type
260-d Model
260-e RNAprep
260-f LabelType
260-g TimePt
260-h State
260-i Comment

(D) RawData (270)
270-a ExptID
270-b seqFile
270-c Probe
270-d RawValue
270-e LocalBgnd/factor
270-f CorrectedValue

Fig. 8

Score	Journal	Score	Journal
100	Nature	80	Brain
100	Science	80	Journal of Physiology
97	Cell	80	Molecular and Cellular Endocrinology
96	Neuron	80	Molecular and Cellular Neurosciences
96	Nature Structural Biology	78	Genome Research
95	Nature Cell Biology	78	Journal of Clinical Epidemiology
95	Nature Genetics	78	Pharmacology
95	Nature Medicine	75	Genomics
95	Nature Neuroscience	75	Journal of Cellular Biochemistry
94	Journal of Structural Biology	75	Journal of Investigative Dermatology
94	New England Journal of Medicine	75	Journal of Medical Genetics
92	Journal of Biological Chemistry	75	Journal of Medical Virology
92	Journal of Lipid Research	75	Journal of Psychiatric Research
92	Journal of Molecular Biology	75	Nitric Oxide
92	Neuroscience	75	Toxicology
91	Journal of Bacteriology	74	Brain and Development
91	Journal of Cell Biology	73	Cancer Immunology, Immunotherapy
91	Journal of Immunology	70	Brain Injury
91	Journal of Neurochemistry	70	Cytogenetics and Cell Genetics
90	Biochemistry	70	Glia
90	Bioinformatics	70	Journal of Biotechnology
90	Immunology	70	Obesity Research
90	Journal of Clinical Investigation	69	Immunopharmacology
90	Journal of Internal Medicine	68	Cellular Signalling
90	Journal of Medicinal Chemistry	68	Immunology and Cell Biology
90	Journal of Membrane Biology	68	Stem Cells
90	Journal of Neuroscience	68	Stroke
90	Journal of Virology	68	Synapse
90	Molecular Microbiology	67	Immunology Letters
89	Cell Biology	65	Cell Motility and the Cytoskeleton
89	Genetics	65	Gene
89	Journal of Experimental Medicine	65	Hormone Research
89	Journal of Neurobiology	65	Journal of Cognitive Neuroscience
89	Neurology	65	Journal of the Neurological Sciences
88	Infection and Immunity	65	Pharmacological Research
87	Journal of Neuroendocrinology	63	American Journal of Physiology
87	Journal of Neuroimmunology	62	American Journal of Medicine
87	Journal of Neurology	62	Journal of Comparative Neurology
87	Neuroendocrinology	60	American Journal of Human Genetics
86	Neuroscience Letters	60	Cell Proliferation
85	Development	55	Journal of Affective Disorders
85	Proceedings National Academy of Sciences	50	DNA and Cell Biology
85	Immunogenetics	45	Anesthesiology
85	Journal of Cell Science	41	Biochemical Journal
85	Journal of Infectious Disease	40	Anesthesia and Analgesia
85	Journal of Neurophysiology	40	Biochemical Pharmacology
85	Molecular Pharmacology	40	Cell Biochemistry and Function
85	Molecular and Cellular Biology	40	Cell and Tissue Research
84	Brain Research	40	Cerebral Cortex
84	Developmental Biology	40	Experimental Cell Research
84	Developmental Neuroscience	40	Histochemistry and Cell Biology
84	Molecular Biology of the Cell	35	Journal of Dermatological Science
83	Cancer	20	Chemistry and Biology
82	Analytical Biochemistry	20	Genes, Chromosomes and Cancer
82	Journal of Endocrinology	20	International Journal of Biochemistry and Cell Biology
82	Journal of Experimental Biology	20	International Journal of Developmental Neuroscience
82	Journal of Histochemistry and Cytochemistry	10	Archives of Biochemistry and Biophysics
82	Journal of Molecular Endocrinology	10	International Journal of Antimicrobial Agents
82	Molecular Immunology	10	International Journal of Cancer
82	Natural Immunity	10	International Journal of Experimental Pathology
82	Nucleic Acids Research	10	Journal of Antisense Research

Fig. 9